ПАРНАЯ РЕГРЕССИЯ

Экономические данные - количественные характеристики каких-либо экономических объектов или процессов.

Экономические данные (фактор 1, фактор 2, ... , фактор n) формируются под действием множества факторов, не все из которых доступны внешнему контролю.

$$A_2 \leq \text{Неконтролируемые факторы} \leq A_1$$

Неконтролируемые факторы могут принимать случайные значения из некоторого множества значений

Случайность экономических данных (Неконтролируемый фактор 1 , Неконтролируемый фактор 2 , ...)

Обусловливают случайность данных, которые они определяют.

Стохастическая (вероятностная) природа экономических данных обусловливает необходимость применения соответствующих статистических методов для их обработки и анализа.

Изучение действительности показывает, что вариация каждого изучаемого признака находится в тесной связи и взаимодействии с вариацией других признаков, характеризующих исследуемую совокупность единиц. Вариация уровня производительности труда работников предприятий зависит от степени совершенства применяемого оборудования, технологии, организации производства, труда и управления и других самых различных факторов.

При изучении конкретных зависимостей одни признаки выступают в качестве факторов, обусловливающих изменение других признаков. Признаки этой первой группы в дальнейшем будем называть признаками-факторами (факторными признаками); а признаки, которые являются результатом влияния этих факторов, будем называть результативными. Например, при изучении зависимости между производительностью труда рабочих и энерговооруженностью их труда уровень производительности труда является

Регрессионный анализ

Регрессионный анализ предназначен для исследования зависимости исследуемой переменной от различных факторов и отображения их взаимосвязи в форме регрессионной модели.

Целью регрессионного анализа является установление формы зависимости между результативным и одним или несколькими факторными признаками. Для решения этой задачи определяется функция (уравнение) регрессии. В статистике под регрессией понимают величину, которая выражает зависимость среднего значения случайной величины y (результативного признака) от значений случайной величины x (факторного признака). Уравнение регрессии выражает среднюю величину одного признака как функцию другого.

Функция регрессии — это модель (уравнение) вида $y = f(x)$, выражающая зависимость переменной y от определяющего ее независимого фактора x.

Парная регрессия

В зависимости от количества факторов, включенных в уравнение регрессии, принято различать простую (парную) и множественную регрессии.

Парная регрессия представляет собой регрессию между двумя переменными – *y* и *x* , т. е. модель вида:

$$y = \hat{f}(x) + \varepsilon$$

где *y* – зависимая переменная (результативный признак); *x* – независимая, или объясняющая, переменная (признак-фактор). Знак «^» означает, что между переменными *x* и *y* нет строгой функциональной зависимости, поэтому практически в каждом отдельном случае величина *y* складывается из двух слагаемых:

$$y = \hat{y}(x) + \varepsilon \qquad (1)$$

где *y* – фактическое значение результативного признака; – теоретическое значение результативного признака, найденное исходя из уравнения регрессии; ε – случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии

Случайная величина ε называется также возмущением. Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

Причин существования случайной составляющей несколько.

1. Не включение объясняющих переменных.

2. Выборочный характер исходных данных

3. Неправильная функциональная спецификация.

4. Возможные ошибки измерения.

Виды регрессий

Различают линейные и нелинейные регрессии. .

Линейная регрессия описывается уравнением

$$\hat{y}_x = a + bx + \varepsilon$$

.

Нелинейные регрессии:

а) степенная ,

$$\hat{y}_x = ax^b$$

в) показательная ,

$$\hat{y}_x = ab^x$$

с) гиперболическая ,

$$\hat{y}_x = a + b/x$$

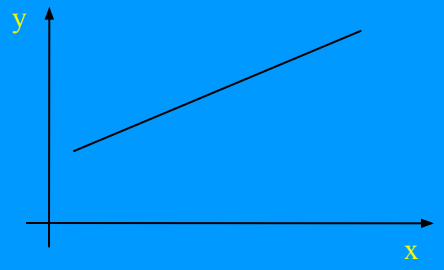
d) параболическая ,

$$\hat{y}_x = a + bx + cx^2$$

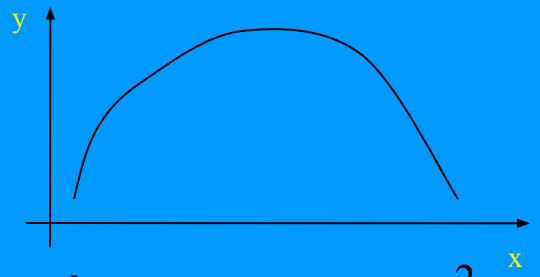
e) полиномы разных степеней .

$$\hat{y}_x = a + bx + cx^2 + dx^3$$

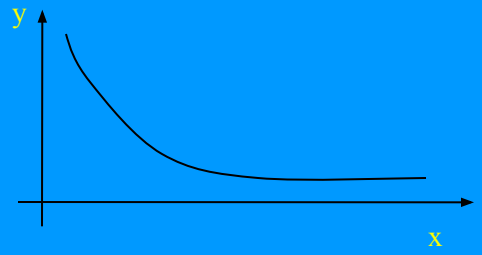
Основные типы кривых, используемые при количественной оценке связей между двумя переменными



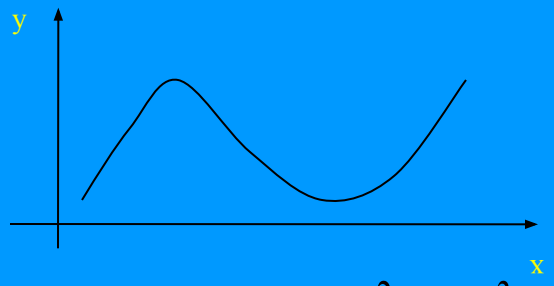
$$\hat{y}_x = a + bx$$



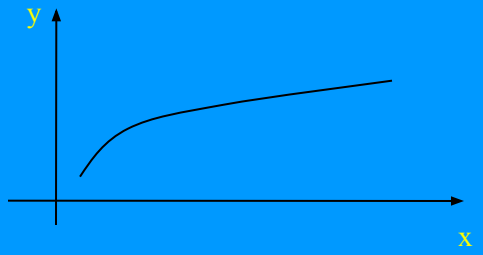
$$\hat{y}_x = a + bx + cx^2$$



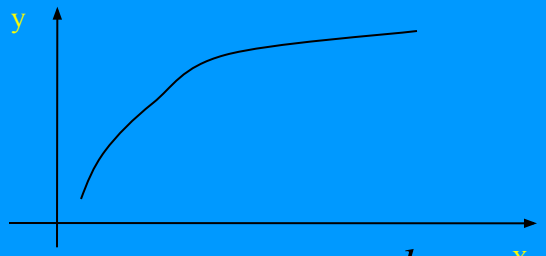
$$\hat{y}_x = a + \frac{b}{x}$$



$$\hat{y}_x = a + bx + cx^2 + dx^3$$



$$\hat{y}_x = ab^x$$



$$\hat{y}_x = ax^b$$

Основные типы кривых, используемые при количественной оценке связей между двумя переменными

$$\hat{y}_x = \frac{1}{a + bx},$$

$$\hat{y}_x = a + bx + \frac{c}{x},$$

$$\hat{y}_x = a + b \lg x,$$

$$\hat{y}_x = \frac{1}{a + bx + cx^2}.$$

$$\hat{y}_x = ax^{-b}$$

$$\hat{y}_x = a + \frac{b}{x}$$

Значительный интерес представляет аналитический метод выбора типа уравнения регрессии, который основан на изучении материальной природы связи исследуемых признаков.

Пусть, например, изучается потребность предприятия в электроэнергии y в зависимости от объема выпускаемой продукции x.

Общее потребление электроэнергии y можно подразделить на две части:

- не связанное с производством продукции a;

- непосредственно связанное с объемом выпускаемой продукции, пропорционально возрастающее с увеличением объема выпуска (bx).

Тогда зависимость потребления электроэнергии от объема продукции можно выразить уравнением регрессии вида.

$$y = a+bx$$

Если разделим обе части уравнения на величину объема выпускаемой продукции (x), то получим выражение зависимости удельного расхода электроэнергии на единицу продукции ($z = y/x$) от объема выпущенной продукции (x) в виде уравнения гиперболы:

$$z = b+a/x$$

Если уравнение регрессии проходит через все точки корреляционного поля, что возможно только при функциональной связи, когда все точки лежат на линии регрессии , то фактические значения результативного признака совпадают с теоретическими $y =$, т.е. они полностью обусловлены влиянием фактора x . В этом случае остаточная дисперсия $\sigma^2_{ост} = 0$.

В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловлено влиянием прочих, не учитываемых в уравнении регрессии, факторов. Иными словами, имеют место отклонения фактических данных от теоретических (y -), где y – фактические значения результативного признака, – расчетные значения, полученные по уравнению регрессии

Величина этих отклонений и лежит в основе расчета остаточной дисперсии:

$$\sigma^2_{ост} = \frac{1}{n}\sum\left(y - \hat{y}_x\right)^2 \qquad (2)$$

Чем меньше величина остаточной дисперсии, тем меньше влияние не учитываемых в уравнении регрессии факторов и тем лучше уравнение регрессии подходит к исходным данным.

Линейная модель парной регрессии и корреляции

Рассмотрим простейшую модель парной регрессии – линейную регрессию. Линейная регрессия находит широкое применение в эконометрике ввиду четкой экономической интерпретации ее параметров.

Линейная регрессия сводится к нахождению уравнения вида

$$\hat{y}_x = a + bx \text{ или } y = a + bx + \varepsilon . \qquad (3)$$

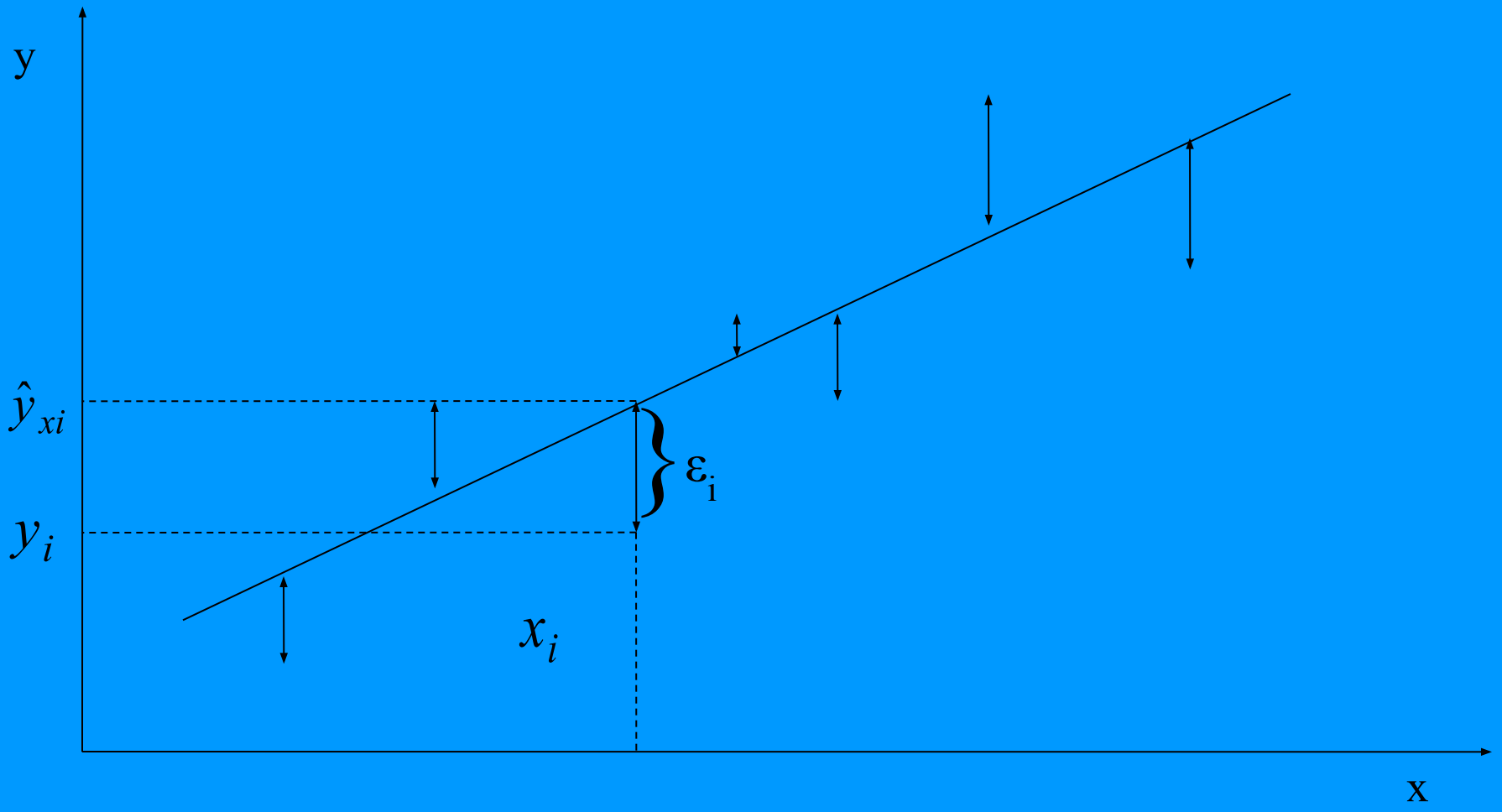
Уравнение вида $\hat{y}_x = a + bx$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x .На графике эти теоретические значения представляют линию регрессии.

Построение линейной регрессии сводится к оценке ее параметров – a и b. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических минимальна:

$$\sum_{i=1}^{n}(y_i - \hat{y}_{xi})^2 = \sum_{i=1}^{n}\varepsilon_i{}^2 \to \min \qquad (4)$$

Т.е. из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной

Линия регрессии с минимальной дисперсией остатков.



Как известно из курса математического анализа, чтобы найти минимум функции (4), надо вычислить частные производные по каждому из параметров a и b и приравнять их к нулю.

Обозначим $\sum \varepsilon_i^2$ через $S(a, b)$, тогда:

$$S(a,b) = \sum (y - a - bx)^2$$

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum (y - a - bx) = 0; \\ \dfrac{\partial S}{\partial b} = -2\sum (y - a - bx) = 0. \end{cases} \qquad (5)$$

После несложных преобразований, получим следующую систему линейных уравнений для оценки параметров a и b:

$$\begin{cases} an + b\sum x = \sum y; \\ a\sum x + b\sum x^2 = \sum xy. \end{cases} \qquad (6)$$

Решая систему уравнений (6), найдем искомые оценки параметров a и b. Можно воспользоваться следующими готовыми формулами, которые следуют непосредственно из решения системы (6):

$$a = \bar{y} - b\bar{x}, \quad b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}, \qquad (7)$$

где $\operatorname{cov}(x, y) = \overline{xy} - \bar{x} \cdot \bar{y}$ - ковариация признаков x и y ,

$\sigma_x^2 = \overline{x^2} - \bar{x}^2$ – дисперсия признака x и

$$\bar{x} = \frac{1}{n}\sum x, \quad \bar{y} = \frac{1}{n}\sum y, \quad \overline{xy} = \frac{1}{n}\sum x \cdot y, \quad \overline{x^2} = \frac{1}{n}\sum x^2.$$

Ковариация – числовая характеристика совместного распределения двух случайных величин, равная математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий.

Дисперсия – характеристика случайной величины, определяемая как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания.

Математическое ожидание – сумма произведений значений случайной величины на соответствующие вероятности.

Параметр b называется коэффициентом регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу. Возможность четкой экономической интерпретации коэффициента регрессии сделала линейное уравнение регрессии достаточно распространенным в эконометрических исследованиях.

Формально a – значение y при x = 0. Если признак-фактор x не может иметь нулевого значения, то вышеуказанная трактовка свободного члена a не имеет смысла, т.е. параметр a может не иметь экономического содержания.

Оценка тесноты связи

Коэффициент корреляции

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает линейный коэффициент корреляции r_{xy}, который можно рассчитать по следующим формулам:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}.$$

Линейный коэффициент корреляции находится в пределах:

$$-1 \le r_{xy} \le 1$$

Чем ближе абсолютное значение r_{xy} к единице, тем сильнее линейная связь между факторами (при $r_{xy} = \pm 1$ имеем строгую функциональную зависимость). Но следует иметь в виду, что близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает отсутствия связи между признаками. При другой (нелинейной) спецификации модели связь между признаками может оказаться достаточно тесной.

Коэффициент детерминации

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции , называемый коэффициентом детерминации. Коэффициент детерминации r_{xy}^2 характеризует долю дисперсии результативного признака y , объясняемую регрессией, в общей дисперсии результативного признака:

$$r_{xy}^2 = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2} \qquad (8)$$

где $\sigma_{ост}^2 = \frac{1}{n}\sum (y - \hat{y}_x)^2; \sigma_y^2 = \frac{1}{n}\sum (y - \bar{y})^2 = \overline{y^2} - \bar{y}^2.$

Соответственно величина $1 - r_{xy}^2$ характеризует долю дисперсии y, вызванную влиянием остальных, не учтенных в модели, факторов.

После того как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

Средняя ошибка аппроксимации

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной. Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации:

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}_x}{y}\right|\cdot 100\%. \qquad (9)$$

Средняя ошибка аппроксимации не должна превышать 8–10%.

Оценка существенности параметров линейной регрессии и корреляции

После того как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

7.1. Оценка значимости уравнения регрессии

Оценка значимости уравнения регрессии в целом производится на основе F -критерия Фишера, которому предшествует дисперсионный анализ. В математической статистике дисперсионный анализ рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели.

7.1.1. Основные положения дисперсионного анализа

Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной y от среднего значения $\bar{y}$ раскладывается на две части – «объясненную» и «необъясненную»:

$$\sum (y-\bar{y})^2 = \sum (\hat{y}_x-\bar{y})^2 + \sum (y-\hat{y}_x)^2$$

где $\sum (y-\bar{y})^2$ – общая сумма квадратов отклонений;

$\sum (\hat{y}_x-\bar{y})^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений);

$\sum (y-\hat{y}_x)^2$ – остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

Общая сумма квадратов отклонений индивидуальных значений результативного признака y от своего среднего значения $\bar{y}$ вызвана влиянием множества причин.

Условно разделим всю совокупность причин на две группы: изучаемый фактор x и прочие факторы. Если фактор не оказывает влияния на результат, то линия регрессии на графике параллельна оси Ox и $y = \bar{y}$.

Тогда вся дисперсия результативного признака обусловлена воздействием прочих факторов и общая сумма квадратов отклонений совпадет с остаточной. Если же прочие факторы не влияют на результат, то y связан с x функционально, и остаточная сумма квадратов равна нулю. В этом случае общая сумма квадратов совпадает с суммой квадратов отклонений, обусловленной регрессией.

Поскольку не все точки поля корреляции лежат на линии регрессии, то всегда имеет место их разброс, как обусловленный влиянием фактора x, т.е. регрессией y по x, так и вызванный действием прочих причин (необъясненная вариация). Пригодность линии регрессии для последующего прогноза зависит от того, какая часть общей вариации признака y приходится на объясненную вариацию. Очевидно, что если сумма квадратов отклонений, обусловленная регрессией, будет много больше остаточной суммы квадратов, то уравнение регрессии статистически значимо и фактор x оказывает существенное воздействие на результат y. Это равносильно тому, что коэффициент детерминации r_{xy} будет приближаться к 1.

7.1.2. Степени свободы

Любая сумма квадратов отклонений связана с числом степеней свободы, т.е. с числом свободы независимого варьирования признака. Число степеней свободы связано с числом единиц совокупности n и с числом определяемых по ней констант m. Применительно к исследуемой проблеме число степеней свободы должно показать, сколько независимых отклонений из *n* возможных

$$\left[(y_1 - \bar{y}), (y_2 - \bar{y}), \ldots, (y_n - \bar{y})\right]$$

требуется для образования данной суммы квадратов, m – число параметров при переменной x

7.1.2.1. Число степеней свободы для общей суммы квадратов

Для общей суммы квадратов

$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

требуется (n - 1) независимых отклонений, ибо по совокупности из n единиц после расчета среднего уровня свободно варьируют лишь (n - 1) - число отклонений. Например, имеем ряд значений *y:* 1, 2, 3, 4, 5. Среднее значение равно 3 и тогда и n отклонений от среднего составят: —2; —1; 0; 1; 2. Так как

$$\sum(y_i - \bar{y}) = 0,$$

то свободно варьируют лишь 4 отклонения, а пятое может быть определено, если предыдущие 4 известны.

7.1.2.2. Число степеней свободы для факторной суммы квадратов

Для факторной суммы квадратов число степеней свободы определяется числом констант при *x*. Для линейной регрессии $\hat{y}_x = a + bx$ при *x* находится коэффициент регрессии *b*, т. е. m = 1, для параболической регрессии $\hat{y}_x = a + bx + cx^2$ при x находятся коэффициенты *b* и *c*, т. е. m = 2, для полинома третьей степени $\hat{y}_x = a + bx + cx^2 + dx^3$ при x находятся коэффициенты *b, c, d* т.е. m = 3.

7.1.2.3. Число степеней свободы для остаточной суммы квадратов

Поскольку существует балансное равенство между числом степеней свободы общей, факторной и остаточной сумм квадратов, то число степеней свободы остаточной суммы квадратов при произвольной регрессии составит n – m –1, т.е. n – 1 = m + (n – m – 1).

7.2. F – критерий Фишера

Схема дисперсионного анализа имеет вид, представленный в таблице 1 (n – число наблюдений, m – число параметров при переменной *x*).

Таблица 1.

Компонент дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum (y-\bar{y})^2$	n - 1	$S^2_{общ} = \frac{\sum (y-\bar{y})^2}{\text{n - 1}}$
Факторная	$\sum (\hat{y}_x-\bar{y})^2$	m	$S^2_{факт} = \frac{\sum (\hat{y}_x-\bar{y})^2}{\text{m}}$
Остаточная	$\sum (y-\hat{y}_x)^2$	n –m -1	$S^2_{ост} = \frac{\sum (y-\hat{y}_x)^2}{n-m-1}$

Поделив каждую сумму квадратов на соответствующее ей число степеней свободы, получим средний квадрат отклонений или, что то же самое, дисперсию S^2 на одну степень свободы. Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F -критерия Фишера:

$$F = \frac{S^2_{факт}}{S^2_{ост}} \qquad (10)$$

Для парной линейной регрессии $m = 1$, поэтому

$$F = \frac{S^2_{факт}}{S^2_{ост}} = \frac{\sum (\hat{y}_x - \bar{y})^2}{\sum (y - \hat{y}_x)^2}(n-2) \qquad (11)$$

Нулевая гипотеза дисперсионного анализа гласит, что коэффициент регрессии равен 0: $b = 0$ и, следовательно, фактор x не оказывает влияния на результат у

Если нулевая гипотеза справедлива, то факторная и остаточная дисперсии не отличаются друг от друга. Для опровержения ее необходимо, чтобы факторная дисперсия превышала остаточную в несколько раз. Разработаны (английским статистиком Снедекором) таблицы критических значений F -отношений при разных уровнях существенности нулевой гипотезы и различном числе степеней свободы. Табличное значение F- критерия — это максимальное значение отношения дисперсий, которое может иметь место при случайном их расхождении для данного уровня вероятности наличия нулевой гипотезы. Фактическое значение F - критерия Фишера (10) сравнивается с табличным значением

$$F(\alpha, k_1, k_2)$$

при уровне значимости α и степенях свободы $k_1 = m$ и $k_2 = n - m-1$.

Вычисленное значение *F*-отношения признается достоверным (отличным от 1), если оно больше табличного. В этом случае отбрасывается нулевая гипотеза об отсутствии связи признаков и делается вывод о существенности этой связи.

Если же значение *F*.-критерия окажется меньше табличного, то вероятность нулевой гипотезы выше заданного уровня (например, 0,05) и она не может быть отклонена без серьезного риска сделать неправильный вывод о наличии связи. В этом случае уравнение регрессии считается статистически незначимым.

7.2.1. Связь *F* - критерия с коэффициентом детерминации

Величина *F* -критерия связана с коэффициентом детерминации r_{xy}^2, и ее можно рассчитать по следующей формуле:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2) \qquad (12)$$

7.3. Оценка значимости коэффициента регрессии

В парной линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью по каждому из параметров определяется его стандартная ошибка: m_b и m_a.

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$m_b = \sqrt{\frac{S^2_{ост}}{\sum (x - \bar{x})^2}} = \frac{S_{ост}}{\sigma_ч \sqrt{n}} \qquad (13)$$

где

$S^2_{ост} = \dfrac{\sum (y - \hat{y}_x)^2}{n - 2}$ - остаточная дисперсия на одну степень свободы.

Величина стандартной ошибки совместно с t –распределением Стьюдента при (n – 2) степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета его доверительного интервала.

Для оценки существенности коэффициента регрессии его величина сравнивается с его стандартной ошибкой, т.е. определяется фактическое значение t -критерия Стьюдента: $t_b = \frac{b}{m_b}$,

которое затем сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы (n - 2). Доверительный интервал для коэффициента регрессии определяется как $b \pm t_{табл} \cdot m_b$

Поскольку знак коэффициента регрессии указывает на рост результативного признака y при увеличении признака-фактора x ($b > 0$), уменьшение результативного признака при увеличении признака-фактора ($b < 0$) или его независимость от независимой переменной ($b = 0$), то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например, $-1{,}5 \le b \le 0{,}8$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

7.4. Оценка значимости коэффициента *a*

Стандартная ошибка параметра *a* определяется по формуле:

$$m_a = \sqrt{S_{ост}^2 \cdot \frac{\sum x^2}{n \cdot \sum (x - \bar{x})^2}} = S_{ост} \cdot \frac{\sqrt{\sum x^2}}{\sigma_x \cdot n} \qquad (14)$$

Процедура оценивания существенности данного параметра не отличается от рассмотренной выше для коэффициента регрессии.

Вычисляется *t* -критерий: $t_a = \frac{a}{m_a}$

его величина сравнивается с табличным значением при (*n* – 2) степенях свободы.